Counterfactual Explainations for Loan Approval Dataset

Mamidala Shivamani

2303A52344

Batch 38

**Objectives and Methodology**

The main goal of this lab exercise is to analyze how minor and realistic modifications in applicant characteristics can influence loan approval decisions using **counterfactual explanations**. The methodology involves importing and preparing a comprehensive loan approval dataset, applying preprocessing steps, training classification models, and generating counterfactual samples through the **DiCE** library. The end-to-end workflow includes data cleaning, encoding of categorical variables, model development (Logistic Regression and Random Forest), and interpretability analysis using counterfactual techniques.

**Dataset Description**

The dataset consists of several thousand entries related to loan applications, with attributes relevant to credit risk evaluation:

**Demographics:** education level, number of dependents.

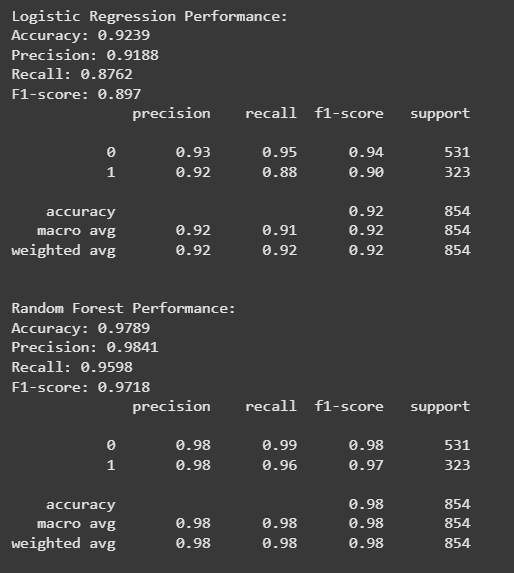
**Financial factors:** applicant’s income, loan amount, loan term, CIBIL score, and various asset values (residential, commercial, luxury, bank).

**Target Variable:** loanstatus (Approved / Rejected).

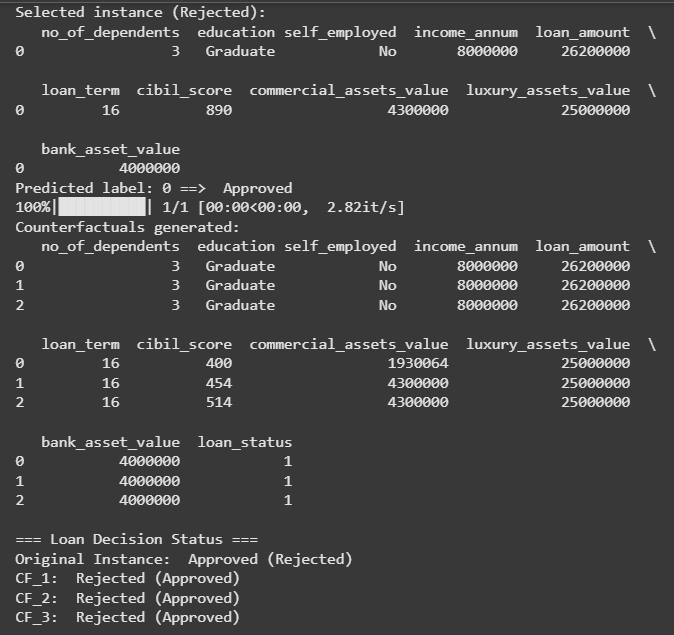
Irrelevant attributes such as “loanid” are removed during preprocessing. Categorical columns (e.g., Graduate/Not Graduate, Self Employed) are encoded appropriately. Missing numerical data is imputed using median values, and missing categorical values are filled with the mode. The dataset shows a balanced distribution of approvals and rejections, making it suitable for effective model training and evaluation.

**Model Performance Results**

Two classification algorithms — **Logistic Regression** and **Random Forest** — are trained using a stratified train-test split. Key performance metrics are evaluated on the test data, such as accuracy, precision, recall, and F1score.



The **Random Forest classifier** exhibits better overall performance and is therefore selected for generating counterfactual explanations. The results confirm good predictive accuracy, while also revealing interesting insights into the model’s sensitivity to different input features.



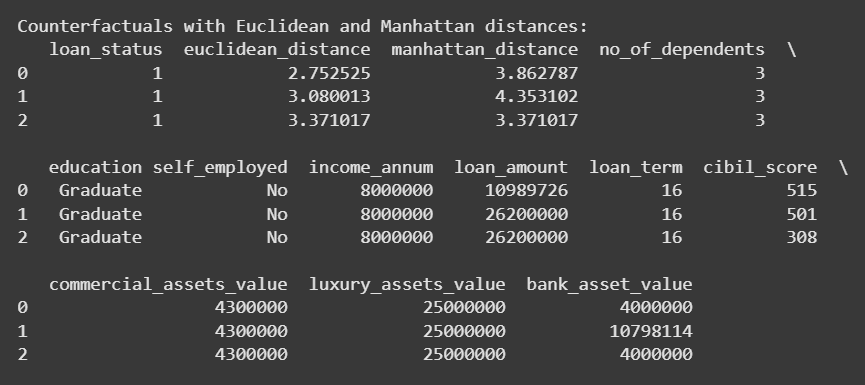
**Counterfactual Examples**

Counterfactual explanations are produced for a few randomly chosen **“Rejected”** applications. Using the DiCE library, feature values are adjusted minimally to flip the model’s prediction from **Rejected** to **Approved**:

Each counterfactual record shows the smallest necessary modifications compared to the original application.

**Distance measures** (Euclidean and Manhattan) are calculated between original and counterfactual points, indicating how close an applicant is to loan approval in the feature space.

These counterfactual examples help visualize exactly which feature adjustments (e.g., income increase, loan amount reduction, or improved credit history) can lead to approval.



**Interpretations and Reflections**

The study demonstrates that **small, practical changes in applicant profiles** can significantly impact loan approval outcomes.

Counterfactual explanations provide clear “what-if” scenarios that improve interpretability for applicants and financial institutions.

They promote **transparency and fairness** by showing actionable paths from rejection to approval.

This approach empowers users by highlighting key factors they can work on to improve their chances in future loan applications.

Overall, applying counterfactual explanations enhances the **trustworthiness and practical utility** of machine learning models in financial decision-making.